PROBABILISTIC GRAPHICAL MODELS

David Madigan

Rutgers University

madigan@stat.rutgers.edu

Expert Systems

- Explosion of interest in "Expert Systems" in the early 1980's

IF the infection is primary-bacteremia
AND the site of the culture is one of the sterile sites
AND the suspected portal of entry is the gastrointestinal tract
THEN there is suggestive evidence (0.7) that infection is bacteroid.

- Many companies (Teknowledge, IntelliCorp, Inference, etc.), many IPO's, much media hype

- Ad-hoc uncertainty handling

Uncertainty in Expert Systems

If A then C (p_1)
If B then C (p_2)

What if both A and B true?

Then C true with CF:

$$p_1 + (p_2 \times (1 - p_1))$$

"Currently fashionable ad-hoc mumbo jumbo"
                                    A.F.M. Smith

Eschewed Probabilistic Approach

- Computationally intractable

- Inscrutable

- Requires vast amounts of data/elicitation

e.g., for n dichotomous variables need $2^n - 1$ probabilities to fully specify the joint distribution

Conditional Independence

$$X \perp\!\!\!\perp Y \mid Z \iff f_{X,Y|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z)$$

Conditional Independence

• Suppose A and B are marginally independent. Pr(A), Pr(B), Pr(C|AB) X 4 = 6 probabilities

• Suppose A and C are conditionally independent given B: Pr(A), Pr(B|A) X 2, Pr(C|B) X 2 = 5

$$A \rightarrow B \rightarrow C$$

$$C \perp\!\!\!\perp A \mid B$$
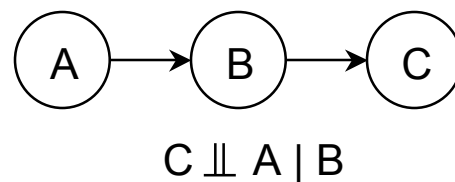
• Chain with 50 variables requires 99 probabilities versus 2^{50}-1

Properties of Conditional Independence (Dawid, 1980)

For *any* probability measure P and random variables A, B, and C:

CI 1: $A \perp\!\!\!\perp B \, [P] \Rightarrow B \perp\!\!\!\perp A \, [P]$

CI 2: $A \perp\!\!\!\perp B \cup C \, [P] \Rightarrow A \perp\!\!\!\perp B \, [P]$

CI 3: $A \perp\!\!\!\perp B \cup C \, [P] \Rightarrow A \perp\!\!\!\perp B \mid C \, [P]$

CI 4: $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C \mid B \, [P] \Rightarrow A \perp\!\!\!\perp B \cup C \, [P]$

Some probability measures also satisfy:

CI 5: $A \perp\!\!\!\perp B \mid C$ and $A \perp\!\!\!\perp C \mid B \, [P] \Rightarrow A \perp\!\!\!\perp B \cup C \, [P]$

CI5 satisfied whenever P has a positive joint probability density with
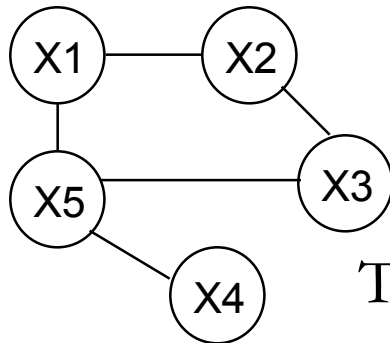respect to some product measure

Markov Properties for Undirected Graphs

(**Global**) S separates A from B $\Rightarrow$ A $\perp\!\!\!\perp$ B | S

(**Local**) $\alpha \perp\!\!\!\perp$ V \ cl(α) | bd (α)

(**Pairwise**) $\alpha \perp\!\!\!\perp \beta$ | V \ {α,β}

$$\boxed{(G) \Rightarrow (L) \Rightarrow (P)}$$

X2 $\perp\!\!\!\perp$ X5, X4 | X1, X3   (1)

$\Rightarrow$ X2 $\perp\!\!\!\perp$ X4 | X1, X3, X5   (2)

To go from (2) to (1) need X5 $\perp\!\!\!\perp$ X2 | X1,X3? or CI5

Lauritzen, Dawid, Larsen & Leimer (1990)

Factorizations

A density f is said to "factorize according to G" if:
$$f(x) = \prod_{C \,\varepsilon\, \mathbf{C}} \psi_C(x_C)$$

"clique potentials"

• cliques are maximally complete subgraphs

Proposition: If f factorizes according to a UG G, then it also obeys the global Markov property

"**Proof**": Let S separate A from B in G and assume $V = A \cup B \cup S$. Let $\mathbf{C}_A$ be the set of cliques with non-empty intersection with A. Since S separates A from B, we must have $B \cap C = \varnothing$ for all C in $\mathbf{C}_A$. Then:
$$f(x) = \prod_{C \in \mathbf{C}_A} \psi_C(x_C) \prod_{C \in \mathbf{C} \backslash \mathbf{C}_A} \psi_C(x_C) = f_1(x_{A \cup S}) f_2(x_{B \cup S})$$

Markov Properties for Acyclic Directed Graphs
(Bayesian Networks)

(**Global**) S separates A from B in $G_{an(A,B,S)}{}^{m} \Rightarrow A \perp\!\!\!\perp B \mid S$

(**Local**) $\alpha \perp\!\!\!\perp nd(\alpha) \backslash pa(\alpha) \mid pa(\alpha)$

$$(G) \Leftrightarrow (L)$$
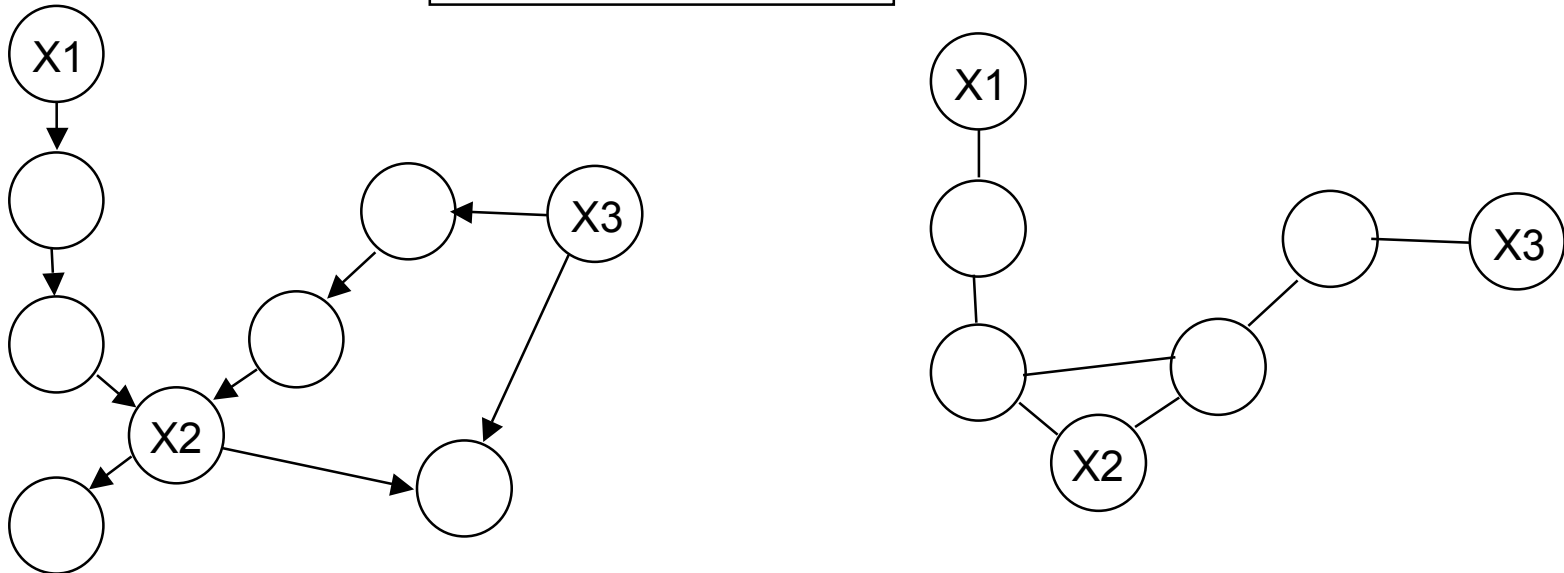
Lauritzen, Dawid, Larsen & Leimer (1990)

Factorizations

A density f admits a "recursive factorization" according to an ADG G if $f(x) = \prod f(x_v \mid x_{\text{pa}(v)})$

$$\boxed{\text{ADG Global Markov Property} \Leftrightarrow f(x) = \prod_{v \, \varepsilon \, V} f(x_v \mid x_{\text{pa}(v)})}$$

Lemma: If P admits a recursive factorization according to an ADG G, then P factorizes according G^M (and chordal supergraphs of G^M)

Lemma: If P admits a recursive factorization according to an ADG G, and A is an ancestral set in G, then P_A admits a recursive factorization according to the subgraph G_A

Factorizations
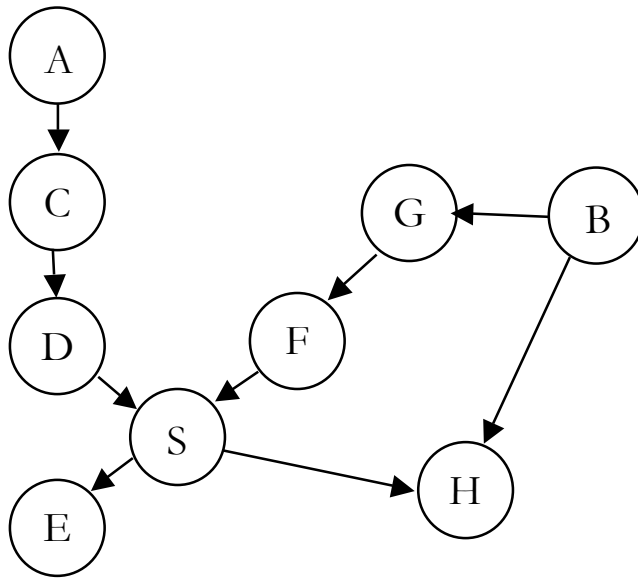
$p(A,B,C,D,E,F,G,H,S) =$

$p(A)p(C|A)p(D|C)p(S|D,F)p(E|S)$
$p(F|G)p(G|B)p(H|S,B)p(B)$

$\Rightarrow$

$p(S|A,B,C,D,E,F,G,H) \propto$
$p(S|D,F)p(E|S)p(H|S,B)$

{*D,F,W,H,B*} is the "*Markov Blanket*" of *S*. It contains the parents of *S*, the children of *S*, and the other parents of the children of *S*.

Markov Properties for Acyclic Directed Graphs
(Bayesian Networks)

(**Global**) S separates A from B in $G_{an(A,B,S)}{}^m \Rightarrow A \perp\!\!\!\perp B \mid S$

(**Local**) $\alpha \perp\!\!\!\perp nd(\alpha)\backslash pa(\alpha) \mid pa(\alpha)$

$\boxed{(G) \Rightarrow (L)}$ $\alpha \cup nd(\alpha)$ is an ancestral set; $pa(\alpha)$ obviously separates α from $nd(\alpha)\backslash pa(\alpha)$ in $G_{an(\alpha\cup nd(\alpha))}{}^m$
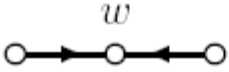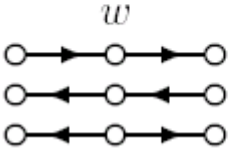
$\boxed{(L) \Rightarrow (factorization)}$ induction on the number of vertices

d-separation

A chain π from a to b in an acyclic directed graph G is said to be *blocked by* S if it contains a vertex $\gamma \in \pi$ such that either:

- $\gamma \in S$ and arrows of π do not meet head to head at γ, or
- $\gamma \notin S$ nor has γ any descendents in S, and arrows of π do meet head to head at g

Two subsets A and B are d-separated by S if all chains from A to B are blocked by S

$w \in \pi^\circ \equiv \pi \setminus \{a, b\}$	$w \in S$	$w \notin S$	
		$w \in \mathrm{an}_D(S)$	$w \notin \mathrm{an}_D(S)$
Either w is a head-to-head node in π:	w is ACTIVE	ACTIVE	BLOCKING
or w is not a head-to-head node in π:	w is BLOCKING	ACTIVE	ACTIVE

d-separation and global markov property

Let A, B, and S be disjoint subsets of a directed, acyclic graph, G. Then S d-separates A from B if and only if S separates A from B in $G_{an(A,B,S)}{}^m$

UG – ADG Intersection
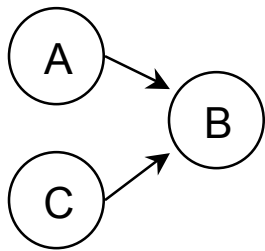
$A \perp\!\!\!\perp C$
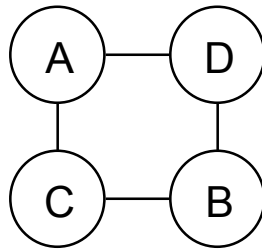
$A \perp\!\!\!\perp B \mid C,D$
$C \perp\!\!\!\perp D \mid A,B$
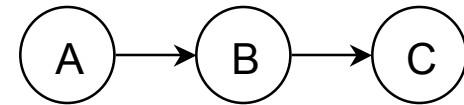
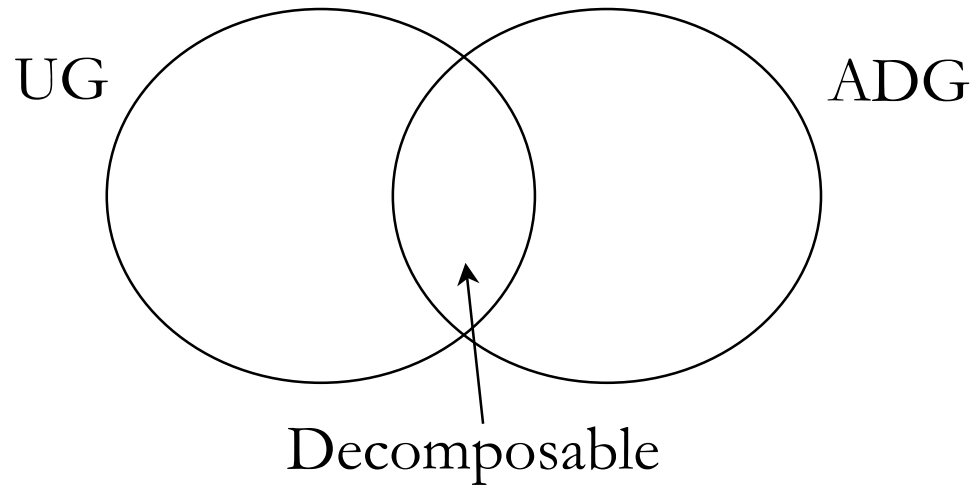$C \perp\!\!\!\perp A \mid B$
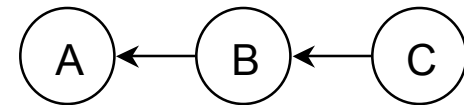
$A \perp\!\!\!\perp C \mid B$
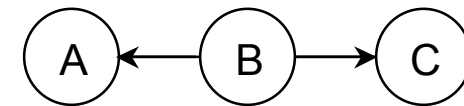
$A \perp\!\!\!\perp C \mid B$
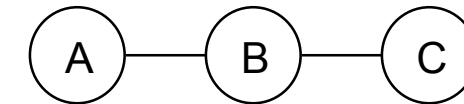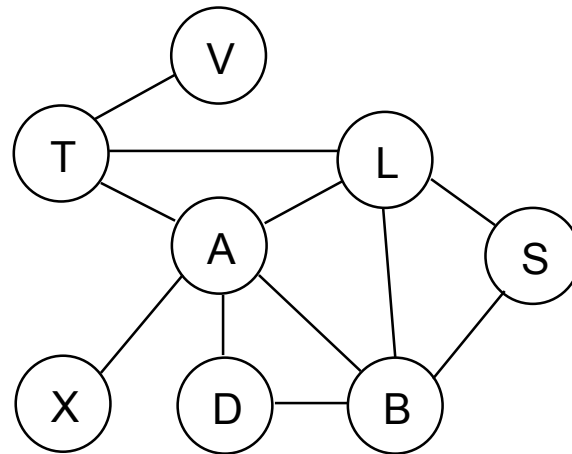
$A \perp\!\!\!\perp C \mid B$

UG – ADG Intersection

UG ⬭⬭ ADG

Decomposable

- UG is decomposable if chordal

- ADG is decomposable if moral

No CI5

- Decomposable ~ closed-form log-linear models

Chordal Graphs and RIP

• Chordal graphs (uniquely) admit clique orderings that have the *Running Intersection Property*

1. {V,T}
2. {A,L,T}
3. {L,A,B}
4. {S,L,B}
5. {A,B,D}
6. {A,X}

• The intersection of each set with those earlier in the list is fully contained in previous set

• Can compute cond. probabilities (e.g. Pr(X|V)) by message passing (Lauritzen & Spiegelhalter, Dawid, Jensen)

Probabilistic Expert System

- Computationally intractable

- Inscrutable

- Requires vast amounts of data/elicitation

- Chordal UG models facilitate fast inference

- ADG models better for expert system applications – more natural to specify $Pr(v \mid pa(v))$

Factorizations

UG Global Markov Property $\Leftrightarrow f(x) = \prod_{C \varepsilon \, C} \psi_C(x_C)$

ADG Global Markov Property $\Leftrightarrow f(x) = \prod_{v \varepsilon \, V} f(x_v \mid x_{\mathrm{pa}(v)})$

Lauritzen-Spiegelhalter Algorithm

$\psi\,(C,S,D) \leftarrow \text{Pr}(S \mid C, D)$
$\psi(A,E) \quad\;\; \leftarrow \text{Pr}(E \mid A)\,\text{Pr}(A)$
$\psi\,(C,E) \quad \leftarrow \text{Pr}(C \mid E)$
$\psi(F,D,B) \leftarrow \text{Pr}(D \mid F)\text{Pr}(B \mid F)\text{Pr}(F)$
$\psi\,(D,B,S) \leftarrow 1$
$\psi\,(B,S,G) \leftarrow \text{Pr}(G \mid S,B)$
$\psi\,(H,S) \quad \leftarrow \text{Pr}(H \mid S)$

- Moralize
- Triangulate
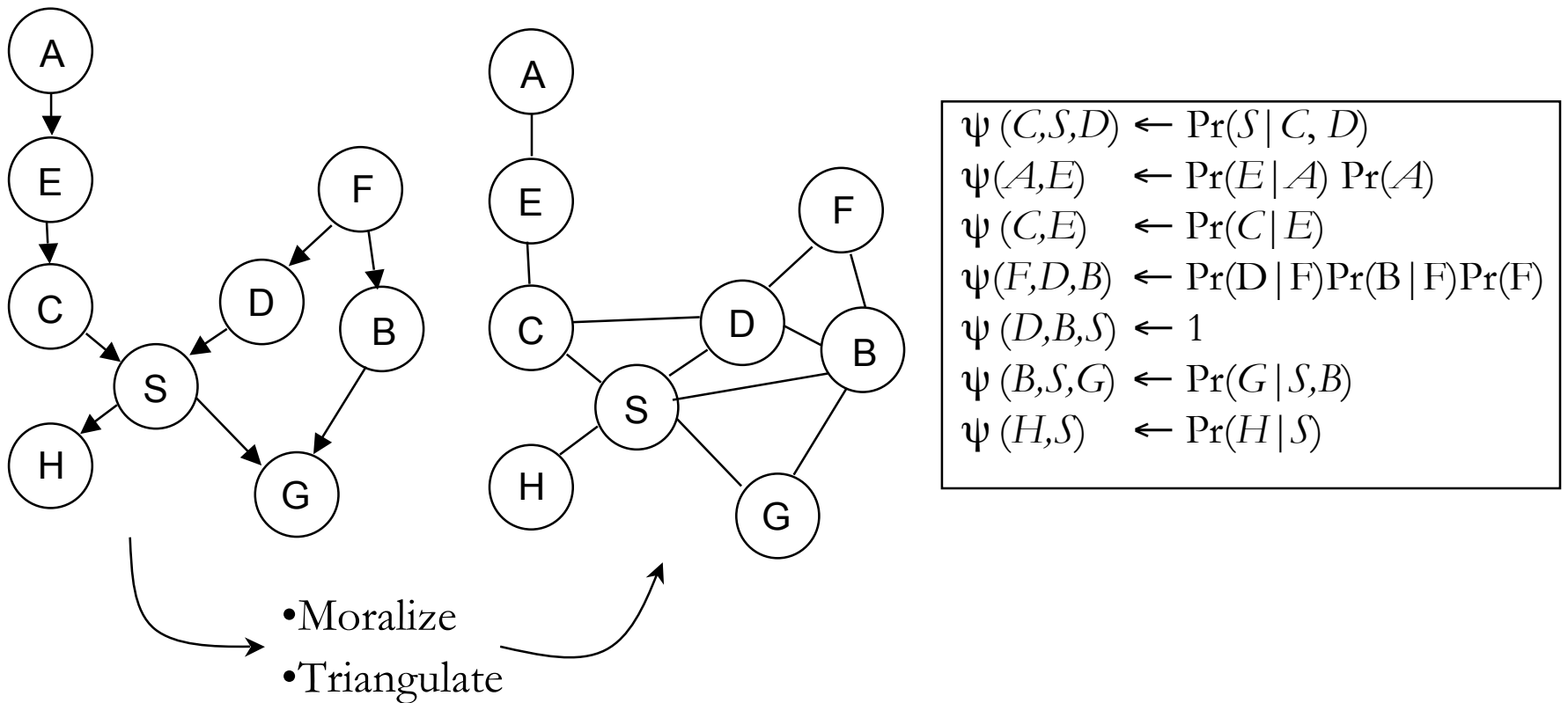
Algorithm is widely deployed in commercial software

L&S Toy Example

A → B → C

Pr(C|B)=0.2 Pr(C|¬B)=0.6
Pr(B|A)=0.5 Pr(B|¬A)=0.1
Pr(A)=0.7

A — B — C

$\psi(A,B) \leftarrow \Pr(B|A)\Pr(A)$
$\psi(B,C) \leftarrow \Pr(C|B)$

AB — [B] — BC
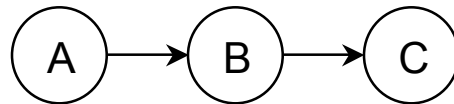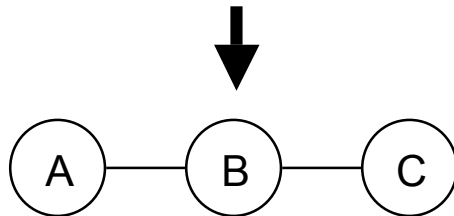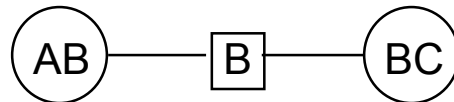
	B	¬B
A	0.35	0.35
¬A	0.03	0.27

	C	¬C
B	0.2	0.8
¬B	0.6	0.4

B	¬B
1	1

Message Schedule: AB→BC   BC→AB

B	¬B
0.38	0.62

	C	¬C
B	0.076	0.304
¬B	0.372	0.248

Pr(A|C)

	C	¬C
B	0.076	0
¬B	0.372	0

Other Theoretical Developments

Do the UG and ADG global Markov properties identify *all* the conditional independences implied by the corresponding factorizations?

Yes. Completeness for ADGs by Geiger and Pearl (1988); for UGs by Frydenberg (1988)

Graphical characterization of collapsibility in hierarchical log-linear models
(Asmussen and Edwards, 1983)

Collapsibility

Clinic A

	Survival		
	No	Yes	
Less	3	176	**1.7%**
More	4	293	**1.4%**

Care

Clinic B

	Survival		
	No	Yes	
Less	17	197	**7.9%**
More	2	23	**8.0%**

Care

Pooled

	Survival		
	No	Yes	
Less	20	373	**5.1%**
More	6	316	**1.9%**

Care

Collapsibility
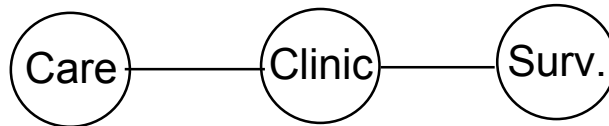
Theorem: A graphical log-linear model L is collapsible onto A iff every connected component of A^c is complete.

Bayesian Learning for Discrete ADG's

- Example: three binary variables
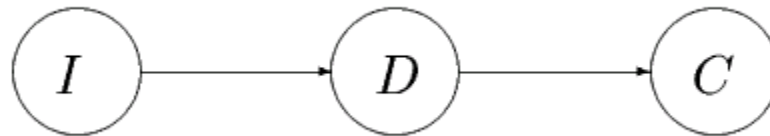
- Five parameters:

$$\mathrm{pr}(C \mid D), \mathrm{pr}(C \mid \overline{D}), \mathrm{pr}(D \mid I), \mathrm{pr}(D \mid \overline{I}) \quad \text{and} \quad \mathrm{pr}(I)$$

Local and Global Independence
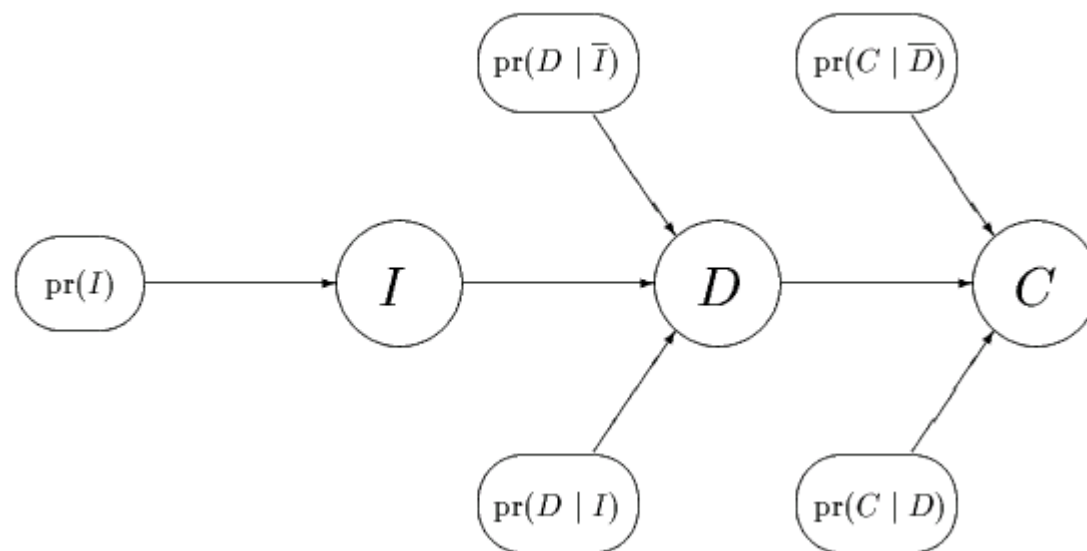
$$\mathrm{pr}(V|\theta) = \prod_{v \in V} \mathrm{pr}(v|\mathrm{pa}(v), \theta_v)$$

$$\theta_C = \{\mathrm{pr}(C \mid D), \mathrm{pr}(C \mid \overline{D})\}, \theta_D = \{\mathrm{pr}(D \mid I), \mathrm{pr}(D \mid \overline{I})\}$$

$$\theta_I = \{\mathrm{pr}(I)\}$$

Bayesian learning

Consider a particular state $pa(v)^+$ of $pa(v)$

$$\text{pr}(v|\text{pa}(v)^+, \theta_v^+) = \theta_v^+$$

We assume that θ_v^+ has a Dirichlet distribution $\mathcal{D}[\lambda_1^+, ..., \lambda_k^+]$

where k is the number of states of v

This prior is conjugate with multinomial sampling, and it follows that

$$\text{pr}(v|\text{pa}(v)^+) = \lambda_v^+ / \sum_i \lambda_i^+$$

If we observe one data case where v is in state j
and the parent state is $\text{pa}(v)^+$, the
posterior distribution of θ_v^+ is given by:

$$\theta_v^+|v \sim \mathcal{D}[\lambda_1^+, ..., \lambda_j^+ + 1, ..., \lambda_k^+]$$

Equivalence Classes and Chain Graphs

- ADG models for a fixed set of vertices decompose into Markov equivalence classes:
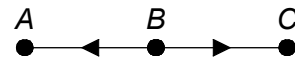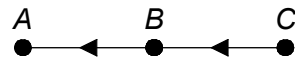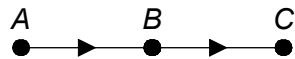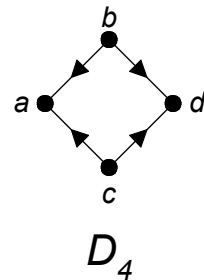

$A \perp\!\!\!\perp C \mid B$
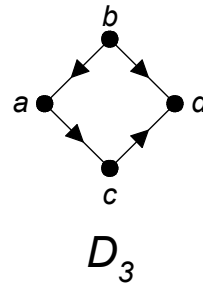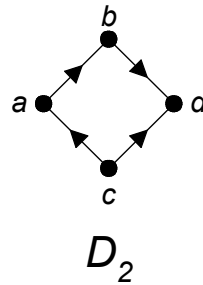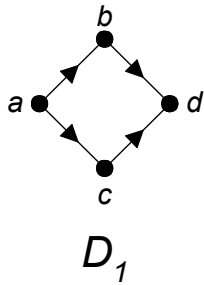
$A \perp\!\!\!\perp D \mid B,C$
$B \perp\!\!\!\perp C \mid A$

$A \perp\!\!\!\perp D \mid B,C$
$B \perp\!\!\!\perp C$

Why is this a problem?

- Repeating analyses for equivalent ADGs leads to significant computational inefficiencies.

- Ensuring that equivalent ADGs have equal posterior probabilities imposes severe constraints on prior distributions (Geiger and Heckerman, 1995).

- Bayesian model averaging procedures that average across ADGs assign weights to statistical models that are proportional to equivalence class sizes.

Equivalence Class Characterization

Theorem (Verma & Pearl, Glymour et al, Frydenberg, AMP94): Two ADGs are Markov equivalent iff they have the same skeletons and the same immoralities.

Definition *The essential graph D* associated with D is the graph*
$$D* := \cup(D'|D' \sim D),$$
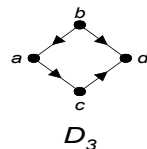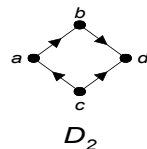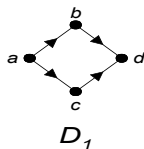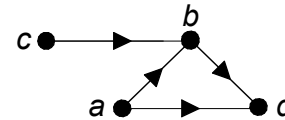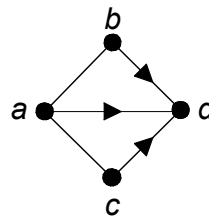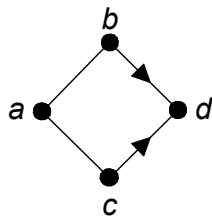
Essential Graphs

AMP (1995)

- Essential graphs are chain graphs

- D* is the unique smallest chain graph Markov equivalent to D

- A graph G = (V, E) is equal to D* for some ADG D if and only if G satisfies the following four conditions:

(i)   G is a chain graph;

(ii)   For every chain component t of G, G_t is chordal;

(iii)   The configuration a→b—c does not occur as an induced

subgraph of G;

(iv)   Every arrow a→b ∈ G is strongly protected in G:

(a)   (b)   (c)   (d)
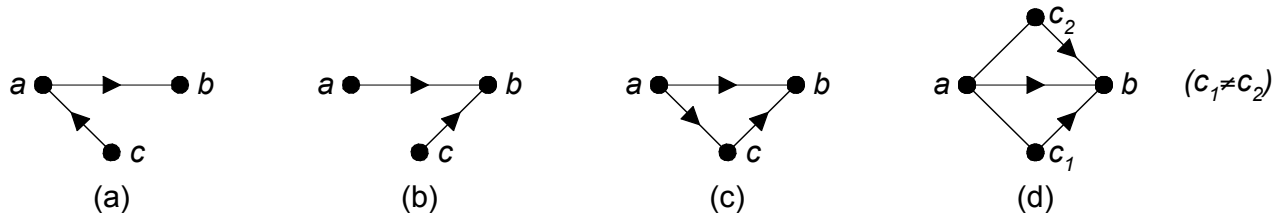
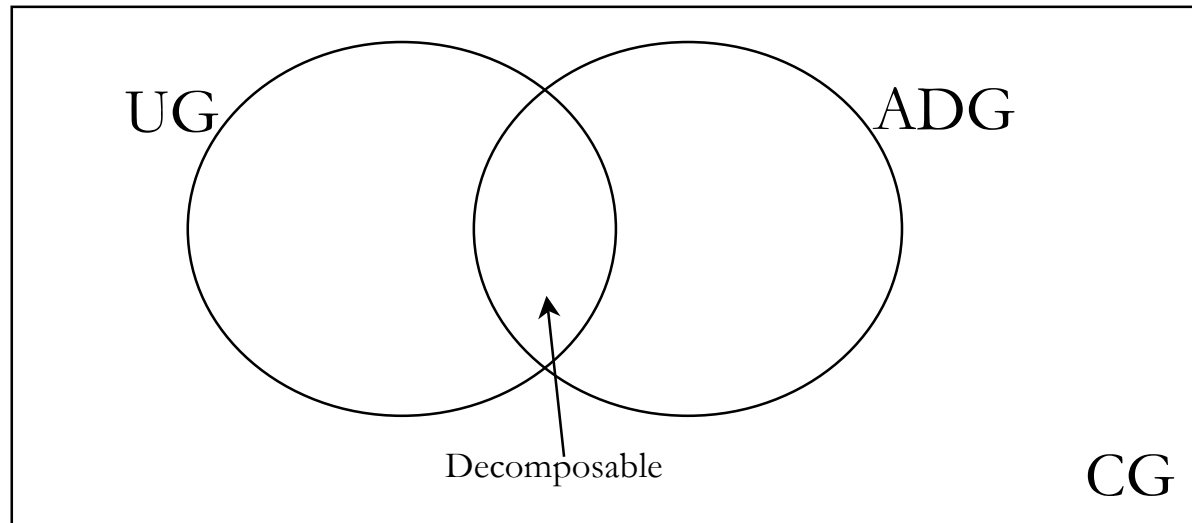also Meek (1995) and Chickering (1995)

What's a Chain Graph?
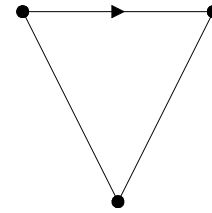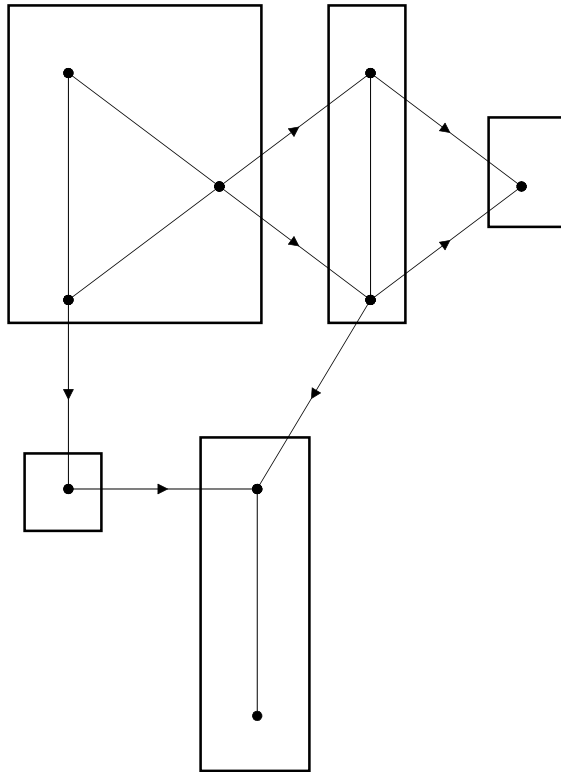
"Equivalence":
$a \sim b$ iff $a \bullet\!\!-\!\!-\!\!\bullet\cdots\bullet\!\!-\!\!-\!\!\bullet\; b$

Chain Components ("Boxes")

Chain Graphs

Decomposable

CG

- Chain graph Markov property, Frydenberg (1990)

- Equivalence results (LWF, AMP, Meek, Studeny)

$C \perp\!\!\!\perp D \mid A,B$ or $C \perp\!\!\!\perp D \mid A$?

Cox & Wermuth (1996)